PREDICTING AFFECTIVE WORD VARIABLES

Hendrik Vankrunkelsven[1], Simon De Deyne[1 2] and Gert Storms[1]
[1] *University of Leuven, Belgium*
[2] *University of Adelaide, Australia*

1. Introduction

Predicting lexical norm data has been done via **text corpora** (e.g., Bestgen & Vincze, 2012; Recchia & Louwerse, 2014; Mandera, Keuleers, & Brysbaert, 2015) and a **word association corpus** (Vankrunkelsven, Verheyen, De Deyne, & Storms, 2015)

We **compare** the quality of prediction using both sources of data.

We predict 3 affective variables: **valence**, **dominance**, and **arousal** as they have been shown to be important in the meaning of words (Osgood, Suci, & Tannenbaum, 1957)

And 2 non-affective variables: **concreteness**, and **age of acquisition** (**AoA**)

We also check whether there is a difference in quality of prediction of **concrete** versus **abstract** words to see if abstract words make use of affective information (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011)

All predictions are cross-validated using lexical norm data

2. Method

2.1. Data

Text corpus:

Syntactic dependency model (De Deyne, Verheyen, & Storms, 2015) :

- Dutch articles in newspapers and magazines
- Internet web pages
- Dutch movie subtitles and Corpus of Spoken Dutch

103,842 lemma types

Similarities (cosine measure)

Word association corpus:

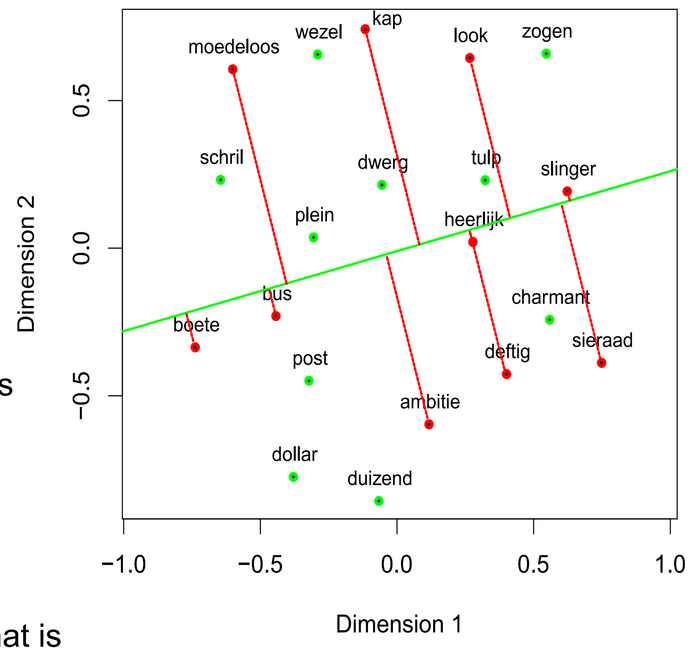
12,566 cue words (De Deyne, Navarro, & Storms, 2013)

Similarities (cosine measure)

2.2. Prediction

MDS-PROFIT:

- Multidimensional scaling (HiT-MDS: 2D - 40D)
- PROFIT: optimal direction in semantic space: multiple linear regression with the variable in question as criterion and the coordinates of the words in the semantic space as predictors
- Prediction word(s): projection(s) on this optimal direction



K-nearest neighbors

Average of norm scores of variable that is predicted from the K-nearest neighbors (based on similarity). K: 1-50, 60, 70, 80, 90, 100

Concrete versus abstract

Median split on concreteness separating the data in relatively concrete and relatively abstract words

2.3. Validation

Cross-validation leave-one-out (L1O)

Prediction using every word except predicted word for the 2831 words available in all datasets (3.1, 3.2, 3.3)

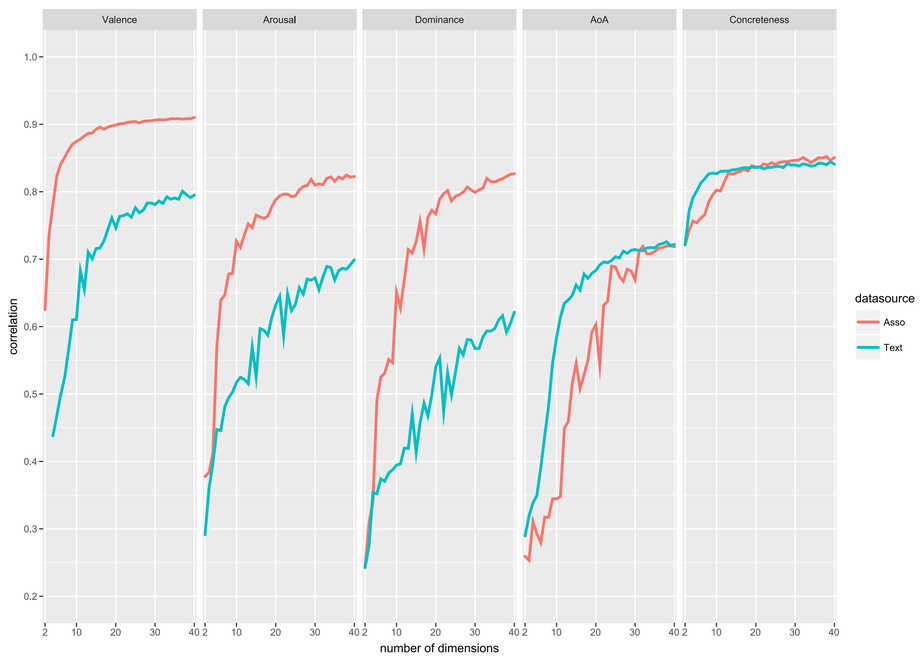
Same procedure but then for relatively concrete and abstract words separately (3.4, 3.5, 3.6)

3. Results

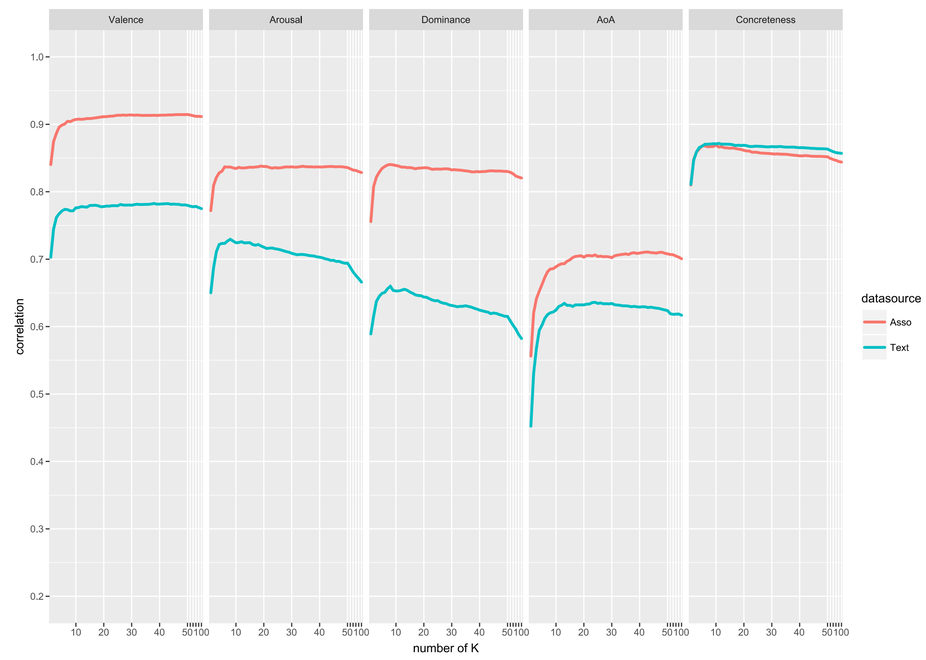
3.1 Best predictions [#dim or #K]

Var:		Val.	Aro.	Dom.	AoA	Con.
PROFIT	Asso.	.91 [40]	.82 [38]	.83 [40]	.72 [39]	.85 [38]
	Text	.80 [37]	.70 [40]	.62 [40]	.73 [38]	.84 [39]
K-near	Asso.	.91 [50]	.84 [19]	.84 [8]	.71 [43]	.87 [10]
	Text	.78 [38]	.73 [8]	.66 [8]	.64 [24]	.87 [11]
K-near W.	Asso.	.92 [50]	.85 [19]	.85 [8]	.73 [48]	.88 [10]
	Text	.79 [43]	.74 [8]	.67 [8]	.64 [24]	.88 [11]

3.2 All predictions PROFIT



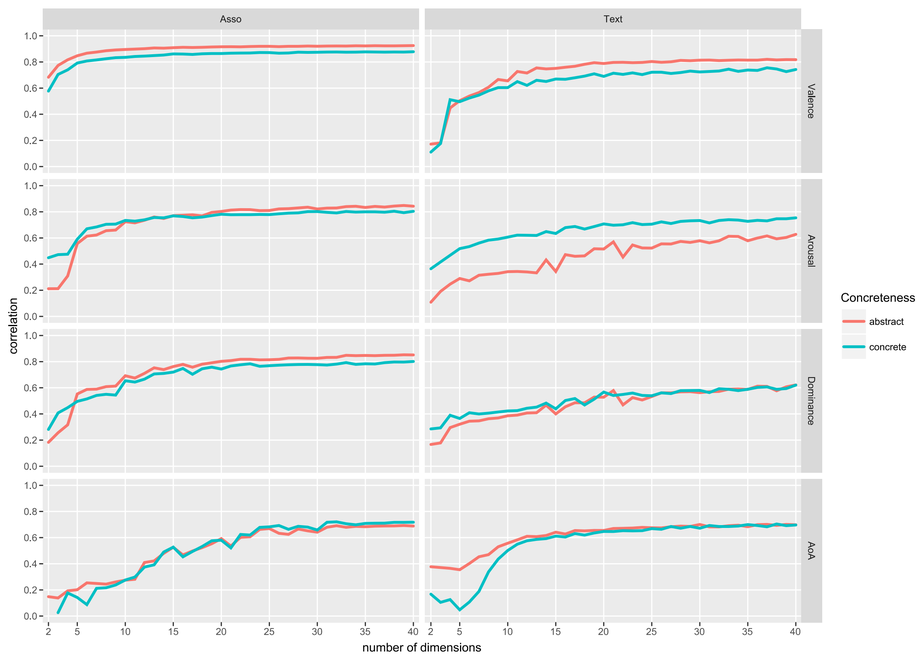
3.3 All predictions K-near



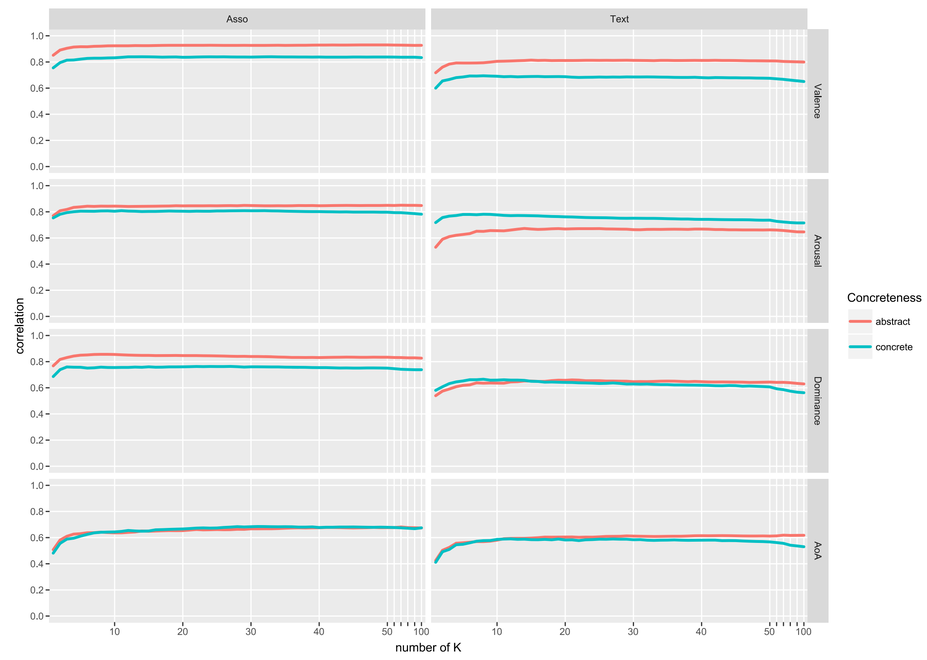
3.4 Best predictions [#dim or #K] concrete vs. abstract

		Asso.		Text	
		Abs.	Con.	Abs.	Con.
PROFIT	Val.	.93 [40]	.88 [40]	.82 [37]	.76 [37]
	Aro.	.85 [39]	.80 [38]	.63 [40]	.75 [40]
	Dom.	.85 [39]	.80 [40]	.62 [40]	.62 [40]
	AoA	.69 [39]	.72 [32]	.70 [39]	.70 [38]
K-near	Val.	.93 [47]	.84 [33]	.81 [15]	.69 [8]
	Aro.	.85 [70]	.81 [32]	.67 [14]	.78 [8]
	Dom.	.86 [9]	.76 [27]	.66 [21]	.67 [8]
	AoA	.68 [70]	.68 [31]	.62 [70]	.59 [12]

3.5 All predictions PROFIT - concrete vs. abstract



3.6 All predictions K-near - concrete vs. abstract



KU LEUVEN

Contact: **hendrik.vankrunkelsven@kuleuven.be**